Generalized Canonical Correlation Analysis for Classification in High Dimensions

Cencheng Shen[a], Ming Sun[a], Minh Tang[a], Carey E. Priebe[a,*]

[a] *Johns Hopkins University, Baltimore, MD 21218-2682*

Abstract

For multiple high-dimensional data sets, we derive conditions under which Generalized Canonical Correlation Analysis improves classification performance, compared to standard Canonical Correlation Analysis using only two data sets. We illustrate our theoretical results with simulations and a real data experiment.

Keywords: Generalized Canonical Correlation Analysis, High-Dimensional Data, Classification, Stiefel Manifold

1. Introduction

Let $(X, Y) \sim F_{XY}$ be an $\mathbb{R}^m \times \{1, \ldots, K\}$ random pair, where X is the feature vector and Y is the class label. In statistical pattern recognition (see, e.g., [6], [7]) one seeks a classifier $g : \mathbb{R}^m \to \{1, \ldots, K\}$ such that the probability of misclassification $L(g) = P\{g(X) \neq Y\}$ is acceptably small. However, in the modern world the feature vector X is generally a random variable of high dimension m, and it is often beneficial to carry out the classification in some lower dimension d ($1 \leq d < m$). Therefore dimension reduction is often applied to first embed X from $\mathbb{R}^m$ to $\mathbb{R}^d$, prior to subsequent classification.

Herein we consider only linear projections, which are commonly used and are the foundation for many nonlinear methods. We denote a linear projection $A : \mathbb{R}^m \to \mathbb{R}^d$ by an $m \times d$ matrix A; then $A'X$ (the $'$ sign denotes transpose) is the projected feature vector in $\mathbb{R}^d$. It follows that

*Corresponding Author: Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682 ; cep@jhu.edu. This work was partially supported by National Security Science and Engineering Faculty Fellowship (NSSEFF), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.

the classification error for a fixed classifier g (whose domain is $\mathbb{R}^d$ from now on) is given by $L_A = P\{g(A^{'}X) \neq Y\}$.

Given a distribution F_{XY}, a classifier g, and a nonempty set of linear projections $\mathcal{A}$, we define an optimal projection $A^* \in \arg\min_{A\in\mathcal{A}}\{L_A\}$ and denote the corresponding minimum error as L_{A^*}. The set $\mathcal{A}$ and the existence of A^* are discussed in Section 2 and Assumption 1. Roughly speaking, L_{A^*} is the minimum error one can hope to achieve by choosing A cleverly among linear projections.

Assuming the classifier g is specified, the crucial problem is how to choose the dimension reduction method. If we have only X available as the feature vector, then PCA (Principal Component Analysis) [1] is a natural choice, which is applied for classification in [15]. On the other hand, if there is an auxiliary feature Z_1 of dimension m_1 available, that is, $(X, Z_1, Y) \sim F_{XZ_1Y}$ on $\mathbb{R}^m \times \mathbb{R}^{m_1} \times \{1, \ldots, K\}$, then CCA (Canonical Correlation Analysis) [10] is applicable on the pair (X, Z_1) to derive the projection A, which is used in [8]. In general, if there are S auxiliary features $\{Z_s \in \mathbb{R}^{m_s}, s = 1, \ldots, S\}$ (we always assume $1 \leq d \leq \min\{m, m_1, \ldots, m_S\}$), then GCCA (Generalized CCA) [11] is applicable on $(X, Z_1, \cdots, Z_S)$ to derive A based on X and the auxiliary features $\{Z_s\}$.

Note that our classification task remains the same, so that at the classification step we observe only X but not $\{Z_s\}$; and so by "GCCA/CCA is applicable" we mean "GCCA/CCA can be used to derive the projection matrix A for use in the classifier $g(A^{'}X)$". Furthermore, although CCA is a special case of GCCA, for clarity purposes we shall assume GCCA uses at least two auxiliary features whenever GCCA is compared to CCA.

In this paper we concentrate our theoretical analysis on GCCA/CCA, and derive conditions implying the superiority of GCCA for classification purposes. Let us say the joint feature $(X, Z_1, \cdots, Z_S) \sim F_{S+1}$, and a projection matrix A derived from GCCA/CCA using X and s auxiliary features is denoted by A_{s+1}. Our main objective is to derive sufficient conditions on F_3 such that if $\max\{L_{A_2}\} = L_{A^*}$, then $L_{A_3} = L_{A^*}$, as well as sufficient conditions such that $L_{A^*} = L_{A_3} < \min\{L_{A_2}\}$. (Note that when there are two auxiliary features, A_2 may come from applying CCA to either (X, Z_1) or (X, Z_2); hence the 'max' and 'min'.) The conditions and necessary prerequisites are discussed in Section 2, and the theorems follow in Section 3. Our theoretical results are illustrated via simulation, as well as a real data experiment on Wikipedia documents, in Section 4. Our previous numerical work illustrating GCCA improvement is available in [12].

2. Preliminaries

Given two auxiliary features Z_1 and Z_2, the joint distribution of (X, Z_1, Z_2) is denoted by $F_3 \in \Omega_3$, where Ω_3 is a family of multivariate distributions on $\mathbb{R}^{(m+m_1+m_2)}$. The overall covariance matrix of F_3 is denoted by

$$\Sigma_{F_3} = \begin{bmatrix} \Sigma_X & \Sigma_{XZ_1} & \Sigma_{XZ_2} \\ \Sigma_{XZ_1}^{'} & \Sigma_{Z_1} & \Sigma_{Z_1Z_2} \\ \Sigma_{XZ_2}^{'} & \Sigma_{Z_1Z_2}^{'} & \Sigma_{Z_2} \end{bmatrix} \in \mathbb{R}^{(m+m_1+m_2)\times(m+m_1+m_2)}.$$

The overall covariance matrix, along with the individual Σ_X, Σ_{Z_1} and Σ_{Z_2}, are all assumed finite and positive semi-definite with rank no less than d.

We can consider GCCA/CCA either with the population covariances or with the sample covariances. For our theoretical analysis we consider the population covariances directly, while in the numerical section we use the sample covariances, which are asymptotically equivalent under standard regularity conditions [1].

Identifying the CCA projection $A_2 = A_2(X, Z_1)$ can be approached as the problem of finding two sets of unit-length canonical vectors $\{a_i\}$ and $\{b_i\}$ to maximize the correlation between $a_i^{'}X$ and $b_i^{'}Z_1$ for each $i = 1, \ldots, d$. (The size of a_i is $m \times 1$ and the size of b_i is $m_1 \times 1$.) That is, we wish to identity

$$\arg\max_{a_i, b_i} \rho_{\{a_i'X, b_i'Z_1\}} = \frac{a_i^{'}\Sigma_{XZ_1}b_i}{\sqrt{a_i'\Sigma_X a_i}\sqrt{b_i'\Sigma_{Z_1}b_i}}, \tag{1}$$

subject to the *uncorrelated constraints*

$$\rho_{\{a_i'X, a_j'X\}} = \frac{a_i^{'}\Sigma_X a_j}{\sqrt{a_i'\Sigma_X a_i}\sqrt{a_j'\Sigma_X a_j}} = 0 \text{ and } \rho_{\{b_i'Z_1, b_j'Z_1\}} = \frac{b_i^{'}\Sigma_{Z_1} b_j}{\sqrt{b_i'\Sigma_{Z_1} b_i}\sqrt{b_j'\Sigma_{Z_1} b_j}} = 0, \forall j < i.$$

Then the $m \times d$ matrix $A_2 = [a_1, \ldots, a_d]$ is the CCA projection matrix for X, and $A_2^{'}X \in \mathbb{R}^d$ is the projected feature vector. Alternatively, a different $A_2 = A_2(X, Z_2)$ can be identified. Note that the arguments to A_2 – (X, Z_1) or (X, Z_2) – represent the choice of auxiliary features, and will be suppressed if the choice is clear or irrelevant in the context.

To identify the GCCA projection A_3 based on (X, Z_1, Z_2), we are looking for three sets of

unit-length canonical vectors $\{a_i\}$, $\{b_i\}$ and $\{c_i\}$ as follows:

$$\begin{aligned} &\arg \max_{a_i, b_i, c_i} \left(\rho^r_{\{a_i'X, b_i'Z_1\}} + \rho^r_{\{b_i'Z_1, c_i'Z_2\}} + \rho^r_{\{a_i'X, c_i'Z_2\}}\right) \\ &\text{subject to } \rho_{\{a_i'X, a_j'X\}} = \rho_{\{b_i'Z_1, b_j'Z_1\}} = \rho_{\{c_i'Z_2, c_j'Z_2\}} = 0, \ \forall j < i. \end{aligned} \tag{2}$$

Then $A_3 = [a_1, \ldots, a_d]$ is the desired GCCA projection. In general, given F_{S+1} we can derive the GCCA projection A_{s+1} for any $1 \leq s \leq S$, and CCA is merely a special case for $s = 1$. The exponent r in the GCCA formulation (2) indicates the specific GCCA criterion, and a common practice is to set $r = 1$ or 2, which maximizes either the sum of correlations or the sum of squared correlations [11]. Because our results are shown to hold for any $r \geq 1$, we implicitly take $r = 1$ unless mentioned otherwise.

Given Σ_X, we shall call an $m \times d$ matrix $A = [a_1, \ldots, a_d]$ a "potential" GCCA projection if and only if its columns $\{a_i\}$ are of unit-length and satisfy the uncorrelated constraints. The set containing all potential GCCA projections is denoted by $\mathcal{A} = \{A |\ \rho_{\{a_i'X, a_j'X\}} = 0\ \forall i \neq j \text{ and } \|a_i\| = 1\ \forall i\}$. As a different choice of auxiliary features yields a different projection, we denote the set containing the GCCA projections A_3 by $\mathcal{A}_3$ and the set containing all CCA projections A_2 by $\mathcal{A}_2$, as well as the set $\mathcal{A}_{s+1}$ in general. Clearly the elements of $\mathcal{A}_{s+1}$ as well as $\mathcal{A}$ depend on Σ_X. Note that the PCA projection is also an element of $\mathcal{A}$, but this is not of any concern in this paper. An important special case: $\mathcal{A}$ represents the Stiefel manifold [3] (containing all orthogonal projections onto dimension d linear subspaces) when Σ_X is a multiple of the identity.

Note that the original GCCA/CCA algorithm does not require the norm of a_i to be the same for all i. We choose them to be unit-length consistently in order to avoid scaling issues in the classification step (alternatively, it is a common practice to set $a_i^{'}\Sigma_X a_i = 1$ for all i, which is equivalent for our purposes). Also note that the choice of the GCCA/CCA projections can be arbitrary. For example, let Σ_X and Σ_{Z_1} be identity matrices and all the singular values of Σ_{XZ_1} be the same; then $A_2(X, Z_1)$ can be chosen arbitrarily in the Stiefel manifold $\mathcal{V}_{d,m}$. In this case A_2 has $md - \frac{d^2+d}{2}$ degrees of freedom, where md comes from the dimension freedom by repeating singular values and $\frac{d^2+d}{2}$ comes from the unit-length requirement and uncorrelated constraints. But if Σ_{XZ_1} does not have repeating singular values, A_2 represents a fixed subspace and has $\frac{d^2-d}{2}$ degrees of freedom, which is implied by the fact that two $m \times d$ matrices A and B represent the same subspace if and only if $AA^{'} = BB^{'}$. The same phenomenon applies for any GCCA projection A_{s+1}.

Returning to the classification problem: given a classifier $g : \mathbb{R}^d \to \{1, \ldots, K\}$ for the low-dimensional feature vector $A^{'}X$, the error L_A may differ for different $A \in \mathcal{A}$. Clearly $\mathcal{A}$ is compact for finite Σ_X and $\{L_A | A \in \mathcal{A}\}$ is bounded between $[0, 1]$, but an optimal low-dimensional projection (with respect to the classification error) is not guaranteed to exist. We make the following assumption to avoid non-existence:

Assumption 1. *Given a classifier g, we assume for the theory in the sequel that an optimal projection $A^* = \arg\min_{A \in \mathcal{A}}\{L_A\}$ exists for any finite Σ_X of rank at least d.*

For example, if the class-conditional distributions $F_{X|Y=k}$ admit probability density functions $f_{X|Y=k}$ for $k = 1, \ldots, K$, then the assumption holds. (In this case L_A is continuous with respect to A, and thus $\{L_A | A \in \mathcal{A}\}$ is compact and admits a minimum.)

By this assumption, the minimum error L_{A^*} always exists and it follows that $L_{A_{s+1}} \geq L_{A^*}$ always holds for any s. Note that the optimal projection A^* need not be unique, since the existence suffices for our purposes. Now we are able to define the notion that GCCA improves CCA using L_{A^*}.

Definition 1. *Assuming the existence of A^*, we say GCCA improves CCA within a family of distributions Ω_3 if and only if $\{F_3 \in \Omega_3 | L_{A_2} = L_{A^*},\ \forall A_2 \in \mathcal{A}_2\} \subset \{F_3 \in \Omega_3 | L_{A_3} = L_{A^*},\ \forall A_3 \in \mathcal{A}_3\}$.*

In general, we say the set of GCCA projections $\mathcal{A}_{s+1}$ improves the set of GCCA projections $\mathcal{A}_{t+1}$ within Ω_{S+1} $(1 \leq s, t \leq S)$ if and only if $\{F_{S+1} \in \Omega_{S+1} | L_{A_{t+1}} = L_{A^*},\ \forall A_{t+1} \in \mathcal{A}_{t+1}\} \subset \{F_{S+1} \in \Omega_{S+1} | L_{A_{s+1}} = L_{A^*},\ \forall A_{s+1} \in \mathcal{A}_{s+1}\}$. (Here the notation "$\subset$" indicates proper subset.)

Put in words, suppose GCCA improves CCA within Ω_3. Then the optimality of both CCA projections implies the optimality of the GCCA projection, and there exists F_3 such that the GCCA projection is optimal while at least one of the CCA projections is not. Note that this is not equivalent to $L_{A_3} \leq L_{A_2}$.

If Ω_3 includes every possible multivariate distribution, then GCCA fails to improve CCA. For example, if Z_1 and Z_2 are both positively correlated to X but Z_1 and Z_2 are negatively correlated, then it might happen that A_2 is optimal while A_3 is not. Hence we look for a family Ω_3 imposing certain relationships among X and $\{Z_s\}$ such that GCCA is guaranteed to improve CCA.

First, we transform X by centering and whitening, so the population mean is zero and the population covariance matrix becomes the identity matrix. Then $\mathcal{A}$ consists of orthogonal projections

onto dimension d linear subspaces, and there exists an orthogonal matrix such that the feature vector can be rotated to guarantee A^* is equivalent to the subspace $\mathbb{R}^d$ spanned by the first d coordinate axes. We denote the transformed random variable by $\tilde{X} = H_X(X - E(X))$, where $E(X)$ is the expectation for centering and H_X is a non-singular $m \times m$ matrix for whitening and rotation. Since the optimal projection for $\tilde{X}$ is spanned by the first d coordinate axes, the form of $\tilde{X}$ based on the class label $Y = \{1, \ldots, K\}$ can be expressed as:

$$\tilde{X} = H_X(X - E(X)) \overset{law}{=} \begin{bmatrix} U_1 I_1 + U_2 I_2 + \cdots + U_K I_K \\ \\ W \end{bmatrix}, \tag{3}$$

where I_k is the class label indicator taking value k with probability p_k and $\sum_{k=1}^{K} p_k = 1$, each $U_k \in \mathbb{R}^d$ is the marginal distribution of $\tilde{X}$ under class k, and $W \in \mathbb{R}^{m-d}$ is the "irrelevant" marginal of $\tilde{X}$. By the above transformation it holds that $E(W) = 0_{(m-d)\times 1}$ and $E(WW^{'}) = I_{(m-d)\times(m-d)}$. Clearly H_X always exists, and there are multiple choices for H_X if A^* is not unique. Now we impose our conditions on F_{S+1} and define what we call the similar family.

Definition 2. *We say the family of distributions Ω^*_{S+1} is* the similar family *if and only if it includes every F_{S+1} such that $(X, Z_1, \cdots, Z_S) \sim F_{S+1}$ satisfies the following conditions:*

Condition (1): For each A^*, there exists non-singular matrices $H_X \in \mathbb{R}^{m\times m}$ and $H_{Z_s} \in \mathbb{R}^{m_s \times m_s}$ for all $s = 1, \ldots, S$, such that Equation (3) holds and there exist non-negative scalars q_{sk} with

$$\tilde{Z}_s = H_{Z_s}(Z_s - E(Z_s)) \overset{law}{=} \begin{bmatrix} q_{s1} U_1 I_1 + q_{s2} U_2 I_2 + \cdots + q_{sK} U_K I_K + e_s \\ \\ W_s \end{bmatrix}, \tag{4}$$

where e_s represents independent noise and $W_s \in \mathbb{R}^{m_s - d}$. Note that unlike H_X, H_{Z_s} need only be non-singular and Z_s are not necessarily whitened and rotated.

Condition (2): $E(U_k U_k^{'}) = I$, and U_k is uncorrelated with W and W_s, for all $k = 1, \ldots, K$ and $s = 1, \ldots, S$.

Condition (3): $\sigma_1(E(W_s W_t^{'})) \leq \sigma_1(E(WW_s^{'}))\sigma_1(E(WW_t^{'}))$ for all $1 \leq s \neq t \leq S$, where we denote $\sigma_i(\Sigma)$ as the ith largest singular value for any matrix Σ henceforth.

Condition (4): $(q_{sk_1} - q_{sk_2})(q_{tk_1} - q_{tk_2}) > 0$ for all $1 \leq s < t \leq S$ and $k_1, k_2 = 1, \ldots, K$; namely

the ordering of coefficients q_{sk} is consistent throughout Z_s.

The purpose of condition (1) is to guarantee that the marginal distribution restricted to A^* of every transformed auxiliary feature under each class is a scalar multiple of the corresponding marginal of $\tilde{X}$ plus error. The possible non-uniqueness of A^* is (mostly) avoided by requiring (1) to hold for any A^*, though the transformation matrices and respective scalars probably differ under different A^*. Condition (2) is to simplify the analysis, without which the proof is much more complex. Given conditions (1) and (2), conditions (3) and (4) are technical conditions used in the proof.

3. Main Results

Theorem 1. *GCCA improves CCA in the similar family Ω_3^*.*

Therefore it is beneficial to use the GCCA projection A_3 within the similar family Ω_3^*. Furthermore, the similar family can be decomposed into three disjoint subsets as follows: $\Omega_3^* = \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} = L_{A_3} = L_{A^*}\} \cup \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A_3} = L_{A^*}\} \cup \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A^*}$ and $L_{A_3} > L_{A^*}\}$, with all the subsets shown to be non-empty and proper in the proof (we can also replace all the 'max' by 'min'). Specifically, if the optimal A^* is known (which may be difficult in practice), then one can check which subset a given $F_3 \in \Omega_3^*$ belongs to according to Inequality (6) and Inequality (7) in the proof below. When the distribution lies in the first or the second subset above, the GCCA projection performs no worse than the CCA projections.

The above theorem can be further generalized to Ω_{S+1}^*.

Corollary 1. *For any $S \geq S' \geq 2$, the set of GCCA projections $\mathcal{A}_{S'+1}$ improves the set of CCA projections $\mathcal{A}_2$ in the similar family Ω_{S+1}^*.*

Under a simplified setting, we can also show that the set of GCCA projections continues to improve when more auxiliary features are included in deriving the projections.

Corollary 2. *Let us replace condition (4) by a simplifying condition (4*): $W_s = W_t$ and $q_{sk} = q_{tk}$ for all $1 \leq s, t \leq S$. Namely the auxiliary features follow the same distribution for $s = 1, \ldots, S$.*

Then for any $S \geq S' \geq 2$, the set of GCCA projections $\mathcal{A}_{S'+1}$ always improves the set of GCCA projections $\mathcal{A}_{S'}$ in the similar family Ω_{S+1}^*.

4. Numerical Experiments

To investigate the performance of the GCCA/CCA projections in classification, we present both numerical simulation and a real data experiment. We use sample covariances to derive the GCCA projections (the algorithm in use is based on [13]) and supervised learning for classification, for which LDA (Linear Discriminant Analysis) [6] is the classification rule.

4.1. Numerical Simulation

We start with four random variables $U_1, U_2 \in \mathbb{R}^3$ and $V_1, V_2 \in \mathbb{R}^6$ all independently normally distributed. The parameters are set as follows: $E(U_1U_1') = E(U_2U_2') = I_{3\times 3}$, $E(U_1) = -E(U_2) = 0.2_{3\times 1}$, $E(V_1V_1') = E(V_2V_2') = 0.5I_{6\times 6}$, $E(V_1) = E(V_2) = 0_{6\times 1}$.

The three random variables $X, Z_1, Z_2 \in \mathbb{R}^9$ are constructed as follows:

$$X \overset{law}{=} \begin{bmatrix} U_1I_1 + U_2I_2 \\ V_1 + V_2 \end{bmatrix}, \quad Z_1 \overset{law}{=} \begin{bmatrix} 0.5U_1I_1 + 0.5U_2I_2 + e_1 \\ V_1 + e_3 \end{bmatrix}, \quad Z_2 \overset{law}{=} \begin{bmatrix} 0.5U_1I_1 + 0.5U_2I_2 + e_2 \\ V_2 + e_4 \end{bmatrix}, \quad (5)$$

where $e_1, e_2 \overset{\text{iid}}{\sim} N(0, 0.75I_{3\times 3})$, $e_3, e_4 \overset{\text{iid}}{\sim} N(0, 0.5I_{6\times 6})$, I_1 and I_2 are class label indicators having equal probability. Using LDA, it is clear that at $d = 3$ the optimal projection A^* uniquely represents the subspace spanned by the first d coordinate axes. Hence we can fit the joint distribution into Definition 2 with $d = 3$, such that $q_{11} = q_{12} = q_{21} = q_{22} = 0.5$, $W = V_1 + V_2$, $W_1 = V_1 + e_3$, $W_2 = V_2 + e_4$, etc. This joint distribution is easily checked to satisfy the required conditions, so it belongs to Ω_3^*. Further, by checking Inequality (6) and Inequality (7) in the proof, the joint distribution is actually an element of the subset $\{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A_3} = L_{A^*}\} \in \Omega_3^*$. So we expect GCCA to outperform CCA when projected onto $\mathbb{R}^3$. Note that in this case we can explicitly calculate L^* for the population model, which is 36.45%.

For each Monte-Carlo replicate, $n = 1500$ observations are generated for each random variable. That is, $\{x^{(1)}, \ldots, x^{(1500)}\}$ for X, $\{z_1^{(1)}, \ldots, z_1^{(1500)}\}$ for Z_1 and $\{z_2^{(1)}, \ldots, z_2^{(1500)}\}$ for Z_2. All data points are used to learn the GCCA/CCA projections A_3/A_2 respectively for $d = 3$. Then the first 1000 points generated from X are projected and used to train the classifier; the remaining 500 points are projected and used for classification error testing. The classification error is recorded separately for the CCA projections $A_2(X, Z_1)$ and $A_2(X, Z_2)$ and for the GCCA projections A_3, using both sum of correlation ($r = 1$) and sum of squared correlation ($r = 2$) criteria. The above is done for 500 Monte Carlo replications, with the average classification error and standard deviation

projections	CCA on (X, Z_1)	CCA on (X, Z_2)	GCCA $(r = 1)$	GCCA $(r = 2)$
average error (L_A)	41.14%	41.33%	37.21%	38.09%
standard deviation	0.15%	0.16%	0.10%	0.12%

Table 1: Simulation Results

topic	category	people	locations	date	math
class label	1	2	3	4	5
article number	119	372	270	191	430

Table 2: Wikipedia Dataset Topics

shown in Table 1 for each projection. The GCCA classification error is lower than CCA as expected, and is fairly close to the optimal error L^*.

4.2. Real Data

The real data experiment applies GCCA/CCA to text document classification. The dataset is obtained from Wikipedia, an open-source multilingual web-based encyclopedia with millions of articles in more than 280 languages. In Wikipedia each article can be related to others in the same language, or articles in other languages with the same subject. Articles of the same subject in different languages are not necessarily exact translations of one another; it is very likely they are written by different people and their contents might differ significantly.

English articles within a 2-neighborhood of the English article "Algebraic Geometry" are collected, and the corresponding French articles of those English documents are also collected, which totals $n = 1382$ pairs of articles in English and French. Let $a_1^e, \ldots, a_{1382}^e$ denote the English articles and $a_1^f, \ldots, a_{1382}^f$ denote the French articles. All articles are manually labeled into 5 disjoint classes $(1 - 5)$ based on their topics, as shown in Table 2.

For the purposes of GCCA/CCA, first we need to embed each article onto the Euclidean space $\mathbb{R}^m$ by Multidimensional Scaling (MDS). MDS [14, 4, 2] strives to give a Euclidean representation while approximately preserving the dissimilarities of the original data: given an $n \times n$ dissimilarity matrix $\Delta = [\delta_{ij}]$ for n observations with δ_{ij} being the dissimilarity measure between the ith and jth observation, MDS generates embeddings $x_i \in \mathbb{R}^m$ for the ith data point to preserve the dissimilarity among the objects pairs, i.e. $||x_i - x_j|| \approx \delta_{ij}$.

For our work two different types of dissimilarity measures are considered for English and French

	Graph Topology Dissimilarity	Text Content Dissimilarity
English articles $\{a_i^e\}$	$\{\bar{x}_i^e\}(GE)$	$\{\hat{x}_i^e\}(TE)$
French articles $\{a_i^f\}$	$\{\bar{x}_i^f\}(GF)$	$\{\hat{x}_i^f\}(TF)$

Table 3: Euclidean Embeddings ($\mathbb{R}^m$) for Wikipedia Articles

articles, giving four dissimilarity matrices of dimension 1382×1382: the graph topology dissimilarity matrix $\bar{\Delta}^e, \bar{\Delta}^f$ and the text content dissimilarity matrix $\hat{\Delta}^e, \hat{\Delta}^f$.

For the graph dissimilarities, $\bar{\Delta}^e$ and $\bar{\Delta}^f$ are constructed based on an undirected graph $G(V, E)$, where V represents the set of vertices of the 1382 Wikipedia documents, and E is the set of edges connecting those articles. There is an edge between two vertices if they are linked in Wikipedia. Then the entry $\bar{\Delta}^e(i, j)$ is calculated from the number of steps on the shortest path from document i to document j in G. For the English articles, $\bar{\Delta}^e(i, j) \in \{0, \ldots, 4\}$, where the 4 comes from the 2-neighborhood document collection. For the French articles $\bar{\Delta}^f(i, j)$ depends on the French graph connections, so it is possible that $\bar{\Delta}^f(i, j) \neq \bar{\Delta}^e(i, j)$. At the extreme end, $\bar{\Delta}^f(i, j) = \infty$ when a_i^f and a_j^f are not connected, and we set $\bar{\Delta}^f(i, j) = 6$ for $\bar{\Delta}^f(i, j) > 4$.

For the text dissimilarities, $\hat{\Delta}^e$ and $\hat{\Delta}^f$ are based on the text processing features for documents $\{a_i^e\}$ and $\{a_i^f\}$. Suppose $\mathbf{z}_i, \mathbf{z}_j$ are the feature vectors for the ith and jth English articles. Then $\hat{\Delta}^e(i, j)$ is calculated by the cosine dissimilarity $\hat{\Delta}^e(i, j) = 1 - \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$. For the experiment we consider the latent semantic indexing (LSI) features [5].

Once different dissimilarity matrices are constructed, the Euclidean space embeddings with $m = 50$ are obtained via MDS. The articles' embeddings are shown in Table 3. At first, English graph dissimilarity (GE) is the classification target, and others (GF, TE, TF) are treated as auxiliary features: all data points are used to learn the GCCA/CCA projections from $\mathbb{R}^m$ to $\mathbb{R}^d$ based on GE and a certain choice of auxiliary features, and the data points of GE are projected by the learned projections. Then 600 observations are randomly picked to train the classifier, with the remaining 782 documents used for classification error testing. We repeat 500 times to calculate the average classification error, for every possible GCCA/CCA projection and various choice of d. The same procedure is repeated with the French graph dissimilarity (GF) being the classification target and the remaining being the auxiliary features. The full results for every possible projection are shown in Figure 1 for the classification of GE. For illustration purposes, two simplified plots are shown in Figure 2 for the classification of GE/GF, for which we omit most projections in order to better

quantify the effects of increasing s (the number of chosen auxiliary features), i.e., only the best A_2 and A_3 are shown. Note that for comparison purposes the PCA projections are also included.

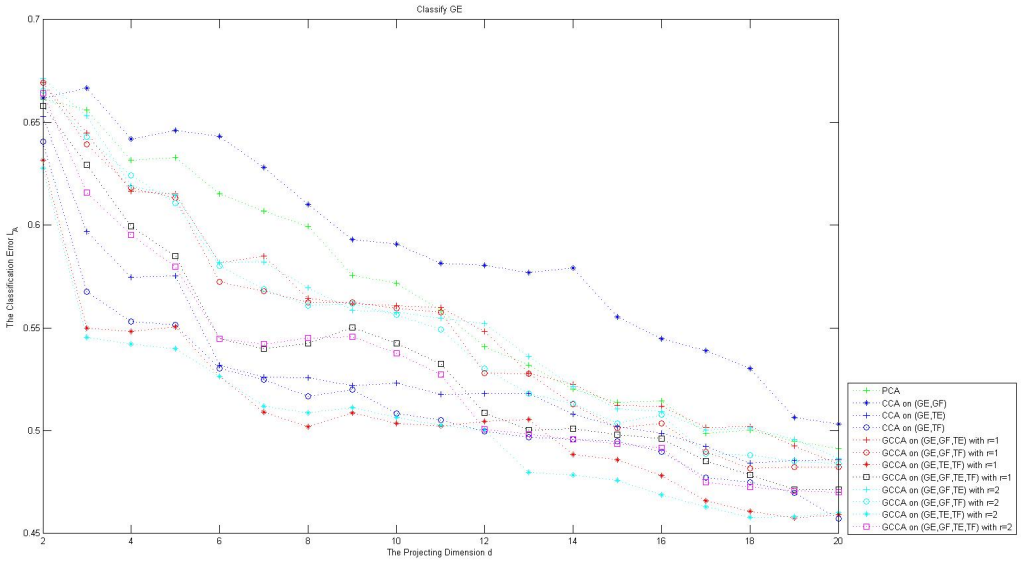
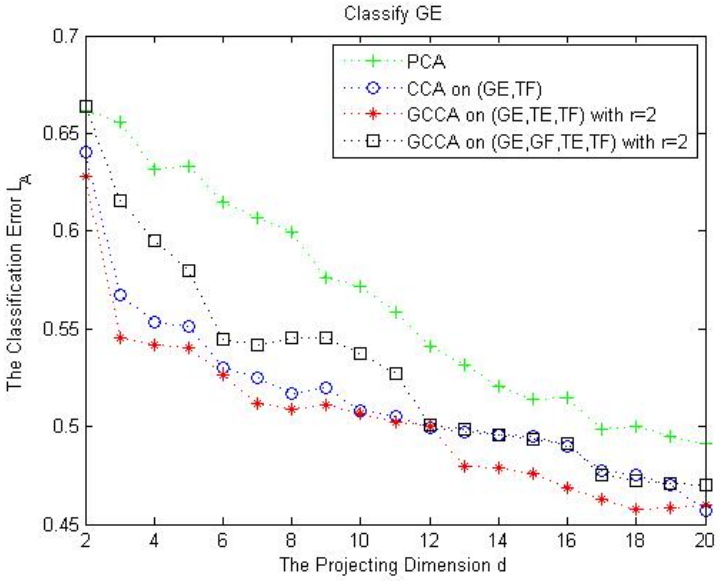


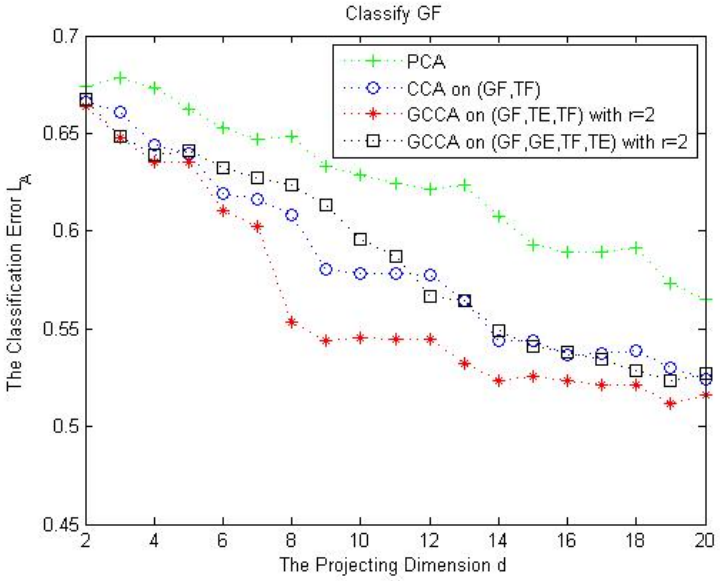
Figure 1: Classification Error for GE

Based on Figure 2, we observe that for most choices of d the best GCCA projection A_3 admits a lower error than the best CCA projection A_2, and both of them are better than the PCA projection. However, it turns out that the GCCA projection A_4 is much worse for classifying the Wikipedia data. This is not a surprise, as one can judge from Figure 1 that the choice of auxiliary features is crucial for the performance of GCCA/CCA projections. From a qualitative perspective, the graph dissimilarities GE and GF are of questionable value because they depend on the Internet links, while the text dissimilarities TE and TF are much more faithful because they are extracted from the document contents. Therefore it is reasonable to believe that choosing a text dissimilarity is better than choosing a graph dissimilarity, which explains why the best A_2 and A_3 do not choose any graph dissimilarity and why A_4 performs worse.

Unfortunately, it is not easy to check the joint distribution by Definition 2, because the optimal projection A^* is unknown. (Even if A^* is known, it is likely the conditions are not satisfied.) Therefore in a real-world application, one must be cautious in adding an auxiliary feature to derive



(a)



(b)

Figure 2: Classification Error for GE/GF (simplified)

the projection, which can be a trial-and-error process.

5. Proofs

5.1. Proof of Theorem 1 when $K = 2$ and $r = 1$

PROOF. We consider $K = 2$ and $r = 1$ here (and generalize in the next proof), so the number of classes is two and the GCCA criterion is the sum of correlations.

If a projection A represents the same subspace as the optimal projection A^* (i.e., $AA^{'} = A^*A^{*'}$), then A is optimal for classification such that $L_A = L_{A^*}$. For most parts it suffices to assume A^* is unique (in the sense of representing the same subspace), which is justified towards the end of the proof.

In addition to the uniqueness of A^*, we also assume that $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices for $s = 1, 2$. This is also justified later, as we will show the theorem is invariant under proper transformations. Further, the expectations $E(X)$ and $E(Z_s)$ are treated as zeros throughout all proofs because the GCCA/CCA projections and the classification task are not affected.

Under the above assumptions, we have the followings: the optimal projection A^* is spanned by the first d coordinate axes; any potential projection $A \in \mathcal{A}$ must be orthonormal and equivalent to an orthogonal projection onto a dimension d linear subspace; and the GCCA/CCA projections A_{s+1} are optimal if and only if $A_{s+1}A_{s+1}^{'} = A^*A^{*'}$.

Because all the pre-multiplication matrices are assumed to be identity matrices, together with conditions (1) and (2) in Definition 2 we have the covariance matrices

$$\begin{aligned}
\Sigma_{XZ_1} &= \begin{bmatrix} pq_{11}E(U_1U_1^{'}) + (1-p)q_{12}E(U_2U_2^{'}) & pE(U_1W_1^{'}) + (1-p)E(U_2W_1^{'}) \\ pq_{11}E(WU_1^{'}) + (1-p)q_{12}E(WU_2^{'}) & E(WW_1^{'}) \end{bmatrix} \\
&= \begin{bmatrix} pq_{11} + (1-p)q_{12} & 0 \\ 0 & E(WW_1^{'}) \end{bmatrix},
\end{aligned}$$

$$\Sigma_{XZ_2} = \begin{bmatrix} pq_{21}E(U_1U_1^{'}) + (1-p)q_{22}E(U_2U_2^{'}) & pE(U_1W_2^{'}) + (1-p)E(U_2W_2^{'}) \\ pq_{21}E(WU_1^{'}) + (1-p)q_{22}E(WU_2^{'}) & E(WW_2^{'}) \end{bmatrix}$$
$$= \begin{bmatrix} pq_{21} + (1-p)q_{22} & 0 \\ 0 & E(WW_2^{'}) \end{bmatrix}.$$

To derive the CCA projection $A_2 = A_2(X, Z_1)$, the two $m \times d$ orthonormal matrices A_2 and B_2 shall maximize the singular values of $A_2^{'}\Sigma_{XZ_1}B_2$ (we take $B_2 = [b_1, \ldots, b_d]$ as in Equation (1), similarly to how we define A_2) [9]. Because A^* represents the dimension d subspace spanned by the first d coordinate axes, $A_2(X, Z_1)$ is optimal if and only if A_2 consists of the first d left singular vectors of Σ_{XZ_1}. Due to the form of Σ_{XZ_1}, in this case B_2 must consist of the first d right singular vectors and the respective correlations are maximized to the decreasingly ordered singular values of the $d \times d$ leading principal sub-matrix of Σ_{XZ_1}. Therefore $A_2A_2^{'} = A^*A^{*'}$ if and only if A_2 is spanned by the first d coordinate axes, or equivalently the largest d singular values of Σ_{XZ_1} all come from the $d \times d$ leading principal sub-matrix.

Putting into inequalities, the CCA projections $A_2(X, Z_s)$ are optimal if and only if,

$$h_s = pq_{s1} + (1-p)q_{s2} - \sigma_1(E(WW_s^{'})) > 0. \tag{6}$$

When either CCA projections is not optimal, at least one h_s is non-positive and represents the "singular value loss" of using CCA.

To derive the GCCA projection A_3 based on (X, Z_1, Z_2), the covariance matrix between Z_1 and Z_2 also comes into play:

$$\Sigma_{Z_1Z_2} = \begin{bmatrix} pq_{11}q_{21}E(U_1U_1^{'}) + (1-p)q_{12}q_{22}E(U_2U_2^{'}) & pq_{11}E(U_1W_2^{'}) + (1-p)q_{12}E(U_2W_2^{'}) \\ pq_{21}E(W_1U_1^{'}) + (1-p)q_{22}E(W_1U_2^{'}) & E(W_1W_2^{'}) \end{bmatrix}$$
$$= \begin{bmatrix} pq_{11}q_{21} + (1-p)q_{12}q_{22} & 0 \\ 0 & E(W_1W_2^{'}) \end{bmatrix}.$$

Argued in a similar manner, the GCCA projection is optimal if and only if A_3 is spanned by the first d coordinate axes. The necessary and sufficient condition for that is

$$h + h_1 + h_2 > 0, \tag{7}$$

where we define $h = pq_{11}q_{21} + (1-p)q_{12}q_{22} - \sigma_1(E(W_1W_2^{'}))$. In words, if both CCA projections are already optimal, it is sufficient that the largest d singular values of $\Sigma_{Z_1Z_2}$ all come from the $d \times d$ leading principal sub-matrix; else if either CCA projections is not optimal, the "singular value gain" from $\Sigma_{Z_1Z_2}$ has to compensate the possible "singular value loss" from Σ_{XZ_1} and Σ_{XZ_2} in order for the GCCA projection to be optimal.

An important step is to prove that if $h_s \geq 0$ for $s = 1, 2$, then $h > 0$. This is true because

$$\begin{aligned}
h &= pq_{11}q_{21} + (1-p)q_{12}q_{22} - \sigma_1(E(W_1W_2^{'})) \\
&\geq pq_{11}q_{21} + (1-p)q_{12}q_{22} - \sigma_1(E(WW_1^{'}))\sigma_1(E(WW_2^{'})) \\
&\geq pq_{11}q_{21} + (1-p)q_{12}q_{22} - (pq_{11} + (1-p)q_{12})(pq_{21} + (1-p)q_{22}) \\
&= p(1-p)(q_{11} - q_{12})(q_{21} - q_{22}) \\
&> 0,
\end{aligned}$$

where the first inequality uses condition (3) in Definition 2, the second inequality is by the fact that $h_s \geq 0$, and the last inequality uses condition (4).

By the above derivation, if both CCA projections are optimal such that $h_s > 0$ for $s = 1, 2$, then Inequality (7) automatically holds and the GCCA projection A_3 is also optimal. This shows that any $F_3 \in \Omega_3^*$ satisfying Inequality (6) for $s = 1, 2$ is an element of the subset $\{F_3 \in \Omega_3^* | \max\{L_{A_2}\} = L_{A_3} = L_{A^*}\}$.

Next we show there exists $F_3 \in \Omega_3^*$ such that Inequality (7) holds while Inequality (6) fails for at least one s. The trivial example is that: if $h_1 = h_2 = 0$, then the GCCA projection is optimal! Furthermore, fixing h, p and all the q_{sk}, the left-hand side of Inequality (7) is clearly continuous with respect to $\sigma_1(E(WW_s^{'}))$ for each s. This means $\sigma_1(E(WW_s^{'}))$ can be increased such that $h_s < 0$ (and condition (3) in Definition 2 will not be violated) while Inequality (7) still holds. So there also exists F_3 such that the GCCA projection is optimal when $h_s < 0$. Thus $\exists F_3 \in \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A_3} = L_{A^*}\}$.

Therefore, when A^* is unique and $H_X, H_{Z_s}, \Sigma_{Z_s}$ all all identity matrices, we proved that: for any given $F_3 \in \Omega_3^*$, if the CCA projections are optimal, so is the GCCA projections; if the CCA projections are not optimal (Inequality (6) is not satisfied for at least one s), the GCCA projection may be optimal (depending on whether the covariance structure satisfies Inequality (7)). Equivalently, we demonstrate that the similarity definition is sufficient for GCCA to improve CCA. Note that the step that ensures $h > 0$ when $h_s \geq 0$ will be used again.

Next we show the result so far is invariant under any $H_X, H_{Z_s}, \Sigma_{Z_s}$ that satisfy Definition 2. Take CCA on (X, Z_1) for an example: by Equation (3) and Equation (4) we have $\Sigma_{\tilde{X}} = H_X \Sigma_X H_X^{'} = I$ and $\Sigma_{\tilde{Z}_1} = H_{Z_1} \Sigma_{Z_1} H_{Z_1}^{'}$; also by eigen-decomposition there exists $m_1 \times m_1$ matrix V s.t. $\Sigma_{\tilde{Z}_1} = V^{'} V$. Then $\Sigma_X = H_X^{-1} H_X^{-1'}$ and $\Sigma_{Z_1} = H_{Z_1}^{-1} V^{'} (H_{Z_1}^{-1} V^{'})^{'}$, and the CCA formulation (1) is equivalent to

$$
\begin{aligned}
& \rho_{\{a_i X, b_i Z_1\}} = \frac{(H_X^{-1'} a_i)^{'} H_X^{'} \Sigma_{X Z_1} H_{Z_1}^{'} V^{-1} (V H_{Z_1}^{-1'}) b_i}{\sqrt{(H_X^{-1'} a_i)^{'} H_X^{-1'} a_i} \sqrt{(V H_{Z_1}^{-1'} b_i)^{'} V H_{Z_1}^{-1'} b_i}}, \\
\text{subject to} \quad & \rho_{\{a_i X, a_j X\}} = \frac{(H_X^{-1'} a_i)^{'} H_X^{-1'} a_j}{\sqrt{(H_X^{-1'} a_i)^{'} H_X^{-1'} a_i} \sqrt{(H_X^{-1'} a_j)^{'} H_X^{-1'} a_j}} = 0 \\
\text{and} \quad & \rho_{\{b_i Z_1, b_j Z_1\}} = \frac{(V H_{Z_1}^{-1'} b_i)^{'} V H_{Z_1}^{-1'} b_j}{\sqrt{(V H_{Z_1}^{-1'} b_i)^{'} V H_{Z_1}^{-1'} b_i} \sqrt{(V H_{Z_1}^{-1'} b_j)^{'} V H_{Z_1}^{-1'} b_j}} = 0,
\end{aligned}
$$

where V^{-1} is defined as the unique Moore-Penrose pseudo inverse if $\Sigma_{\tilde{Z}_1}$ is singular. Hence it is equivalent to consider the projections $H_X^{-1'} A_2$ and $V H_{Z_1}^{-1'} B_2$ on $(\tilde{X}, V^{-1'} \tilde{Z}_1)$ (both $\tilde{X}$ and $V^{-1'} \tilde{Z}_1$ are of identity variance) with covariance $H_X^{'} \Sigma_{X Z_1} H_{Z_1}^{'} V^{-1}$, instead of the projections A_2 and B_2 on (X, Z_1). The same holds for the GCCA formulation (2). Furthermore, the classification task remains the same because the projected feature $A^{'} X = (H_X^{-1'} A)^{'} H_X X$ is invariant under the full-rank transformation H_X. Therefore the optimal projection A^* and the GCCA/CCA projections A_{s+1} are all equivalent to the identity variance case up to H_X, and the result is clearly invariant.

At last we justify the case when A^* is not unique, which means there exists A^* that is spanned by the first d coordinate axes under different transformation matrices. Because the conditions in Definition 2 are required to be satisfied for all A^*, in most cases the CCA optimality is still equivalent to Inequality (6), i.e., CCA is optimal if and only if Inequality (6) is satisfied for at least one A^* after proper transformations for each A^*. The same holds for the GCCA optimality (Inequality (7)), and we can still conclude that GCCA improves CCA following the same steps. However, a

special case should be taken into consideration, and we take the CCA projection $A_2(X, Z_1)$ for an illustration: Suppose the singular vector $\sigma_1(E(WW_s^{'}))$ corresponds to is the $(d+1)$th coordinate axes and $\sigma_1(E(WW_s^{'})) > \sigma_2(E(WW_s^{'}))$. Then $A_2(X, Z_1)$ can be chosen to represent any dimension d subspace of the space spanned by the first $(d+1)$ coordinate axes, and the degrees of freedom is $(d+1)d - \frac{d^2+d}{2}$ (the degrees of freedom may increase if there are repeating singular values). Now, if A^* happens to have the same degrees of freedom in the space spanned by the first $(d+1)$ coordinate axes, then $A_2(X, Z_1)$ is optimal if and only if $h_1 \geq 0$ (rather than $h_1 > 0$) because any arbitrary choice of A_2 is optimal. Similar phenomenon applies for A_{s+1}, in which case Inequality (6) and Inequality (7) should be adjusted to include equalities. However, in this case we still have $h + h_1 + h_2 > 0$ when the CCA projections are optimal, which is still sufficient (but may not be necessary) for GCCA to be optimal. Therefore, GCCA still improves CCA in case of non-unique A^*, and the justification is done. □

5.2. Proof of Theorem 1 for any $K \geq 2$ and $r \geq 1$

PROOF. Now we generalize the result to arbitrary $K \geq 2$ (multi-class) and any $r \geq 1$ (the GCCA criterion). Without loss of generality, we assume A^* is unique and $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices.

Using the setting in Equation (4) and argue similarly as before, GCCA improves CCA if and only if

$$h = \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(W_1 W_2^{'})) > 0 \tag{8}$$

is true when $h_s = \sum_{k=1}^{K} p_k q_{sk} - \sigma_1(E(WW_s^{'})) \geq 0$ for $s = 1, 2$.

This is true because

$$\begin{aligned}
h &= \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(W_1 W_2^{'})) \\
&\geq \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(WW_1^{'}))\sigma_1(E(WW_2^{'})) \\
&\geq \sum_{k=1}^{K} p_k q_{1k} q_{2k} - h_1 h_2 \\
&\geq \sum_{1 \leq k_1 < k_2 \leq K} p_{k_1} p_{k_2} (q_{1k_1} - q_{1k_2})(q_{2k_1} - q_{2k_2}) \\
&> 0,
\end{aligned}$$

where the inequalities again follow from conditions (3) and (4) and simple algebra.

As to the GCCA criterion with $r \geq 1$, GCCA improves CCA if and only if

$$(\sum_{k=1}^{K} p_k q_{1k} q_{2k})^r - \sigma_1^r(E(W_1 W_2^{'})) > 0$$

is true when $h_s \geq 0$. Clearly this inequality holds if and only if it holds for $r = 1$, which is Inequality (8). Hence it is true and GCCA improves CCA in the similar family for any $r \geq 1$.

Thus Theorem 1 is proved for any number of classes and any GCCA criterion with $r \geq 1$. □

5.3. Proof of Corollary 1 and Corollary 2

PROOF. Without loss of generality, we carry out the proof assuming A^* is unique, $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices, and $K = 2$ and $r = 1$.

There are S auxiliary features in total, and thus $\binom{S}{S'}$ choices of auxiliary features for $A_{S'+1}$. We define $h_s = pq_{s1} + (1-p)q_{s2} - \sigma_1(E(WW_s^{'}))$ and $h_{st} = pq_{s1}q_{t1} + (1-p)q_{s2}q_{t2} - \sigma_1(E(W_s W_t^{'}))$ for any s and t satisfying $S \geq s, t \geq 1$, where h_{st} is a generalization of h in the proof of Theorem 1.

Then the GCCA projection $A_{S'+1}$ using the first $S^{'}$ auxiliary features is optimal if and only if

$$\sum_{1 \leq s < t \leq S'} h_{st} + \sum_{s=1}^{S^{'}} h_s > 0. \tag{9}$$

This is a generalization of Inequality (7), because there are $S^{'}$ possible "singular value loss" caused by Σ_{XZ_s} and $\frac{S^{'}(S^{'}-1)}{2}$ additional cross-covariance terms $\Sigma_{Z_s Z_t}$ between the auxiliary features. Note that for any other $A_{S'+1} \in \mathcal{A}_{S'+1}$ with a different choice of auxiliary features, we can still use Inequality (9) for the optimality by switching the first $S^{'}$ auxiliary features with the chosen $S^{'}$ auxiliary features.

All the CCA projections are optimal if and only if $h_s > 0$ for all $s = 1, \ldots, S$. This implies that $h_{st} > 0$ is always true for any $1 \leq s < t \leq S$, and Inequality (9) holds for any $A_{S'+1} \in \mathcal{A}_{S'+1}$ with $S \geq S^{'} \geq 2$. Therefore the set of GCCA projections $\mathcal{A}_{S'+1}$ always improves the set of CCA projections $\mathcal{A}_2$, and Corollary 1 is proved.

To prove Corollary 2, we use the simplifying condition (4*). Then Inequality (9) simplifies to $\frac{S^{'}-1}{2} h_{12} + h_1 > 0$, because h_{st} are the same for all $1 \leq s, t \leq S^{'}$ and so are h_s. We need to show that if $A_{S'}$ are optimal for certain F_{S+1}, so is $A_{S'+1}$. (note that the choice of auxiliary features no

longer matters because they follow the same distribution, which means all the elements in $\mathcal{A}_{S'+1}$ represent the same subspace.)

When $S^{'} = 2$, it is a special case of Theorem 1 because any F_{S+1} satisfying condition (4*) also satisfies condition (4). Clearly A_2 is optimal if and only if $h_1 = h_2 > 0$, which implies $h_{12} > 0$. So Inequality (9) holds and A_3 is also optimal.

When $S^{'} = 3$, A_3 is optimal if and only if $h_{12} + h_1 > 0$. In this case if $h_1 > 0$, then we have $h_{12} > 0$; if $h_1 < 0$, then $h_{12} > 0$ must be true in order for A_3 to be optimal. In any case, $\frac{3}{2}h_{12} + h_1 > 0$ is true and A_4 is optimal.

Therefore, the optimality of A_3 implies the optimality of A_4. By induction, for any $S \geq S^{'} \geq 2$, the optimality of $\mathcal{A}_{S'}$ implies the optimality of $A_{S'+1}$ under the simplifying condition (4*). Corollary 2 is proved. (Unfortunately this is not true under the original condition (4), and one can easily make up a counter-example by checking Inequality (9).) □

References

References

[1] T. W. Anderson. *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics, 3rd edition, 2003.

[2] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications.* Springer-Verlag, 2005.

[3] Y. Chikuse. *Statistics on special manifolds, Lecture Notes in Statistics.* Springer, 2003.

[4] T. Cox and M. Cox. *Multidimensional Scaling.* Chapman and Hall, 2001.

[5] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[6] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley, 2nd edition, 2000.

[8] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[9] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[10] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[11] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[12] M. Sun, C. E. Priebe, and M. Tang. Generalized canonical correlation analysis for disparate data fusion. *Pattern Recognition Letters*, 34(2):194–200, 2013.

[13] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.

[14] W. Torgerson. *Multidimensional scaling: I. theory and method*. Psychometrika, 1952.

[15] J. Yang and J. Y. Yang. Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36(2):563–566, 2003.